

Data Lake

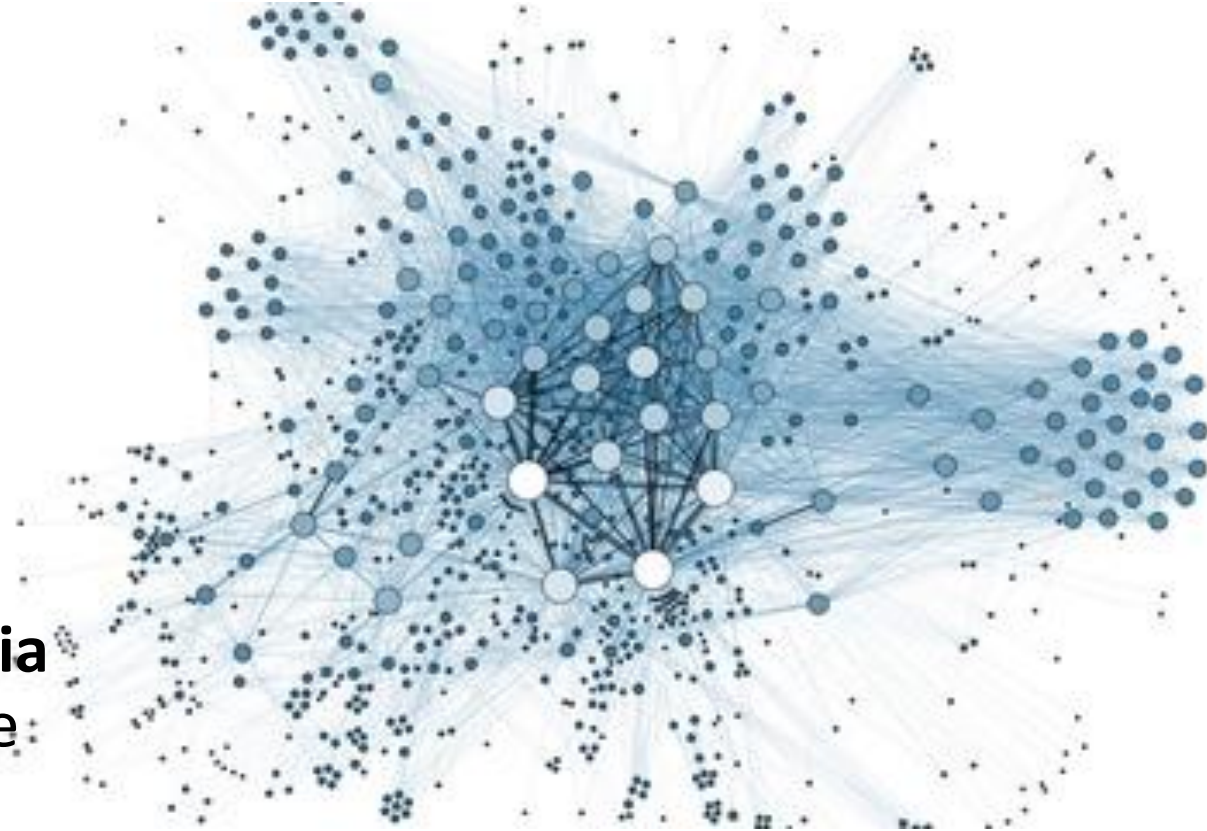
La virtualizzazione dei dati

Fulvio Sbroiavacca



Big Data

Come un microscopio,
i **Big Data** ci consentono di
indagare
i **più piccoli dettagli**
ed al tempo stesso di
“vedere” **correlazioni su ampia
scala**, finora sconosciute, dalle
potenzialità infinite



Big Data analytics: la vera innovazione
deriva proprio dalla capacità di
elaborare enormi quantità crescenti di
informazioni in tempo reale per
utilizzarne i risultati e **generare
conoscenza**

Big Data - Processi

Per definire cosa sono i Big Data pensiamo un attimo al nostro quotidiano: interazioni sui social network, un click su un sito web, una ricerca su Google, un nostro acquisto al supermercato, una foto, un messaggio vocale, un tweet, i nostri smartphone interconnessi ...

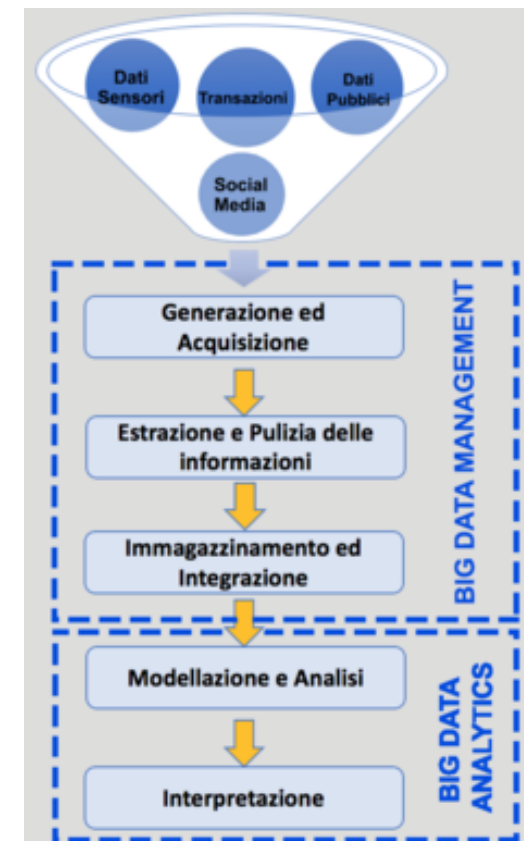
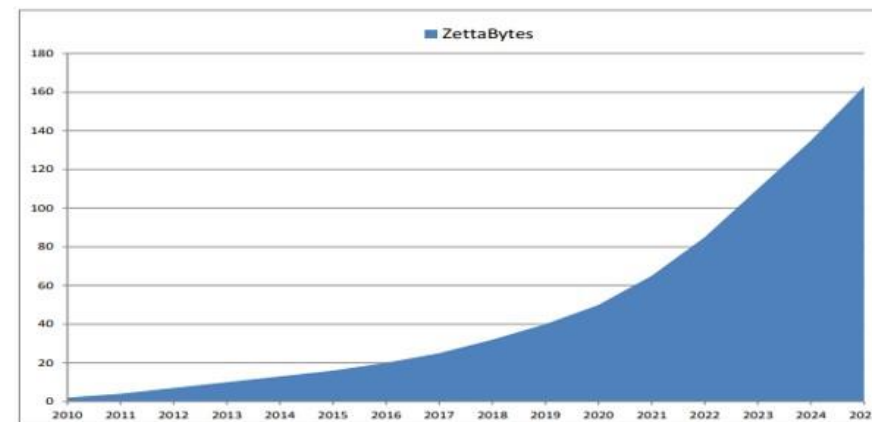
Tutto ciò genera una mole di dati incredibilmente più elevata di qualche decennio fa: enormi volumi di dati eterogenei per fonte e formato, analizzabili in tempo reale

I processi principali che compongono il ciclo di vita dei Big Data

- **Big Data Management**
i processi e le tecnologie per l'acquisizione, la memorizzazione, la preparazione ed il recupero
- **Big Data Analytics**
i processi utilizzati per analizzare e acquisire informazioni utili da grandi dataset allo scopo di interpretare e descrivere il passato (descriptive analytics), predire il futuro (predictive analytics) o consigliare azioni (prescriptive analytics)

https://it.wikipedia.org/wiki/Big_data

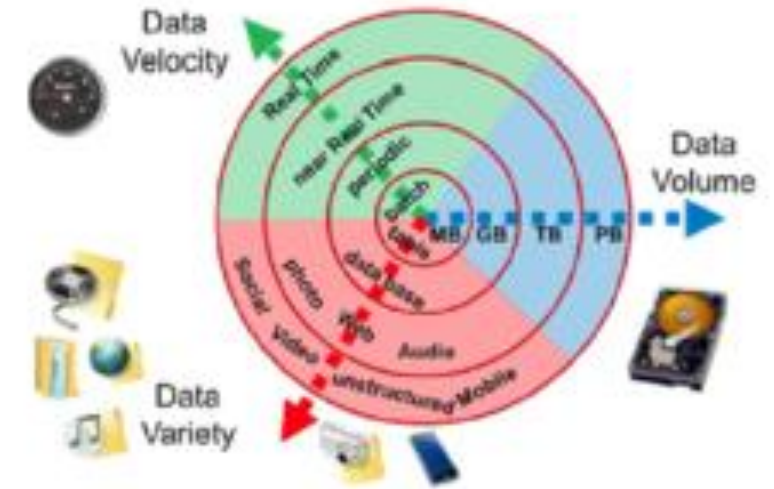
https://blog.osservatori.net/it_it/big-data-cosa-sono



Big Data - Analytics

I Big Data presentano tre caratteristiche fondamentali:

- **Volume**
si tratta di grandissime quantità di dati
- **Velocità**
enorme di produzione e memorizzare le informazioni
- **Varietà**
la provenienza dei dati può essere la più varia



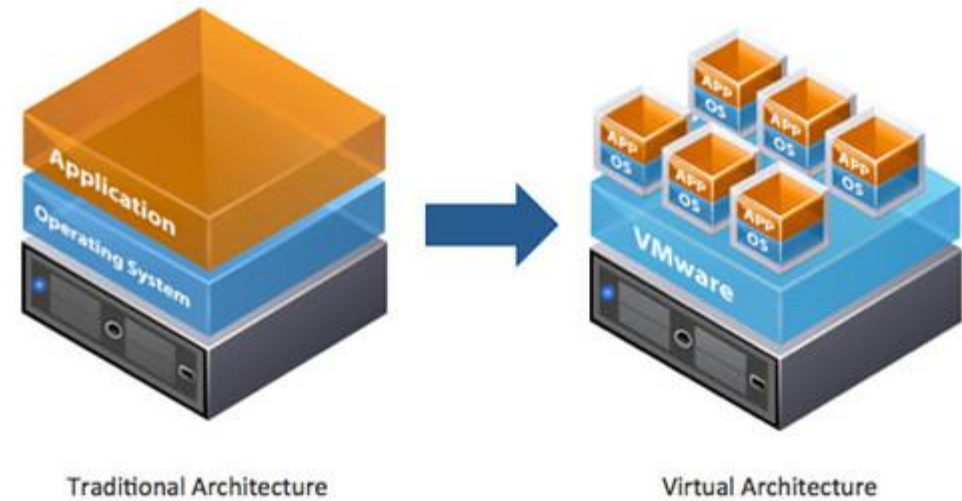
I dati social, ad esempio, hanno la caratteristica di essere a **bassa strutturazione** o **non strutturati** e cioè difficilmente inquadrabili con le tradizionali tecniche di organizzazione dei database
Ad esempio in social come Instagram troviamo foto collegate a hashtag con like e indicazioni geolocalizzate

Le caratteristiche dei Big Data hanno portato all'ideazione dei **data lake** (lago di dati) ed al concetto di **virtualizzazione** dei dati

La virtualizzazione

Che cos'è la virtualizzazione?

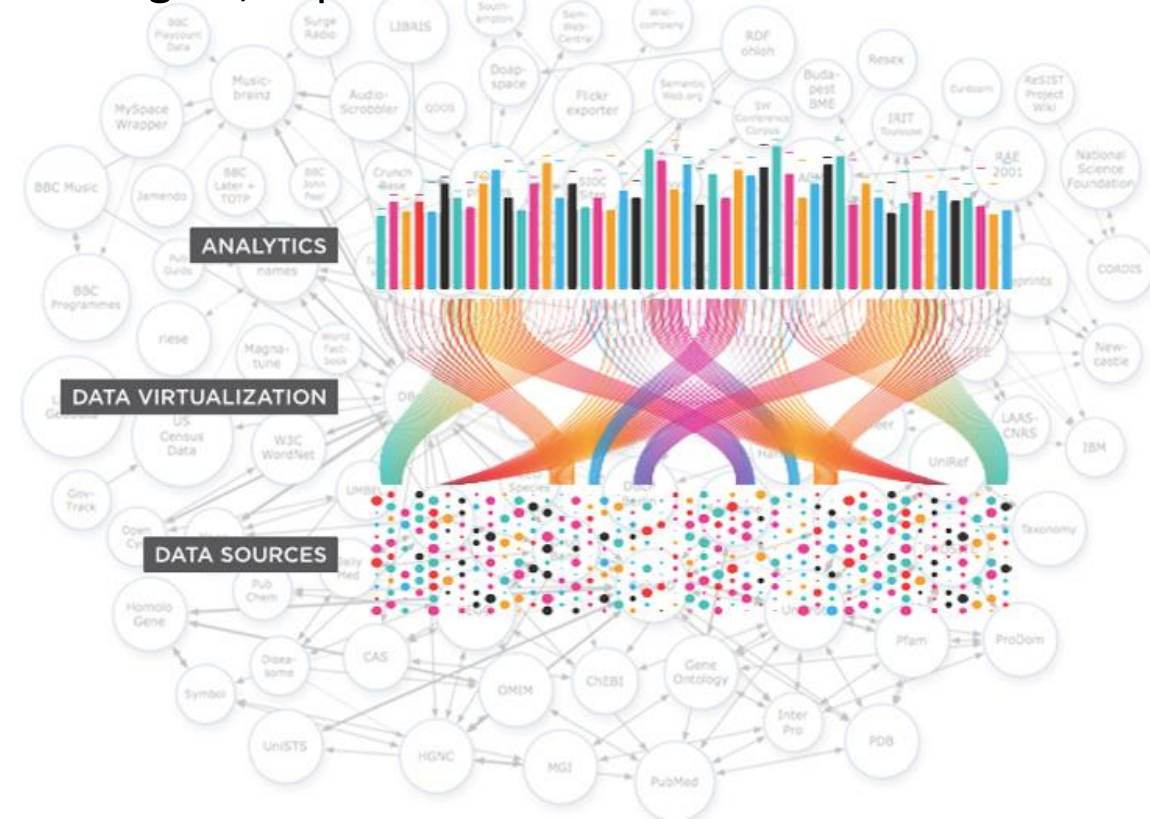
- **La virtualizzazione consiste in un'astrazione dalle risorse fisiche**
- Una componente IT creata con la virtualizzazione si indica come **componente virtuale** o logica e può essere utilizzata esattamente come il suo corrispondente fisico
- Definizione IT: per virtualizzazione si intende l'astrazione di risorse IT fisiche come hardware, software, memoria e componenti di rete, il fine consiste nel migliorare l'utilizzo delle risorse IT, fornendo risorse a livello virtuale distribuendole in modo flessibile a seconda delle esigenze
- La virtualizzazione è una tecnologia informatica che utilizza uno **strato software per simulare macchine fisiche e periferiche** (dischi, schede di rete ecc.) e usarle come se fossero fisiche
- Una macchina virtuale non è altro che un insieme di file, serve poi un software in grado di utilizzare tali file per consentire all'utente di gestire la macchina (avviarla, fermarla, modificare l'hardware, clonarla ecc).



La virtualizzazione dei dati

La virtualizzazione dei dati è un approccio per unificare i dati da più origini in un **singolo livello** in modo che le applicazioni, gli strumenti di reporting e gli utenti finali possano accedere e manipolare i dati **senza richiedere dettagli tecnici** come l'origine, la posizione e le strutture

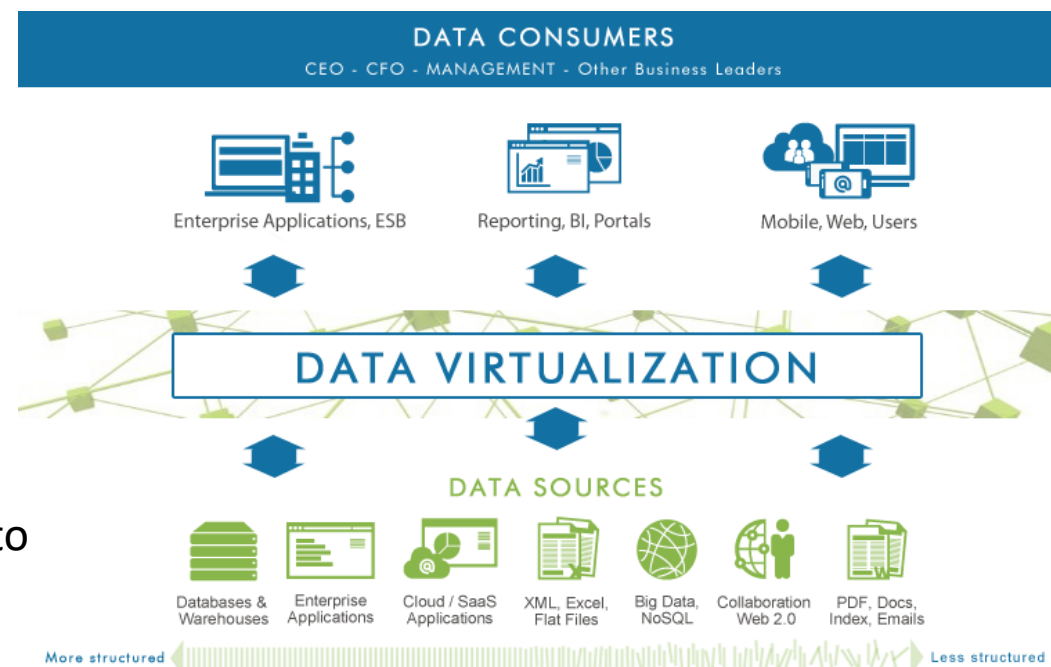
- Viene fornito l'**accesso in tempo reale** al sistema di origine per i dati
- Per risolvere le differenze nei formati e nella semantica tra sorgente e utilizzo (*consumer*) vengono utilizzate varie tecniche di **astrazione e trasformazione**



Data Lake

Un Data Lake è una specie di repository di dati in grado di **virtualizzare** set di dati non elaborati di grandi dimensioni e di varia tipologia, strutturati e non strutturati, nel loro **formato nativo**

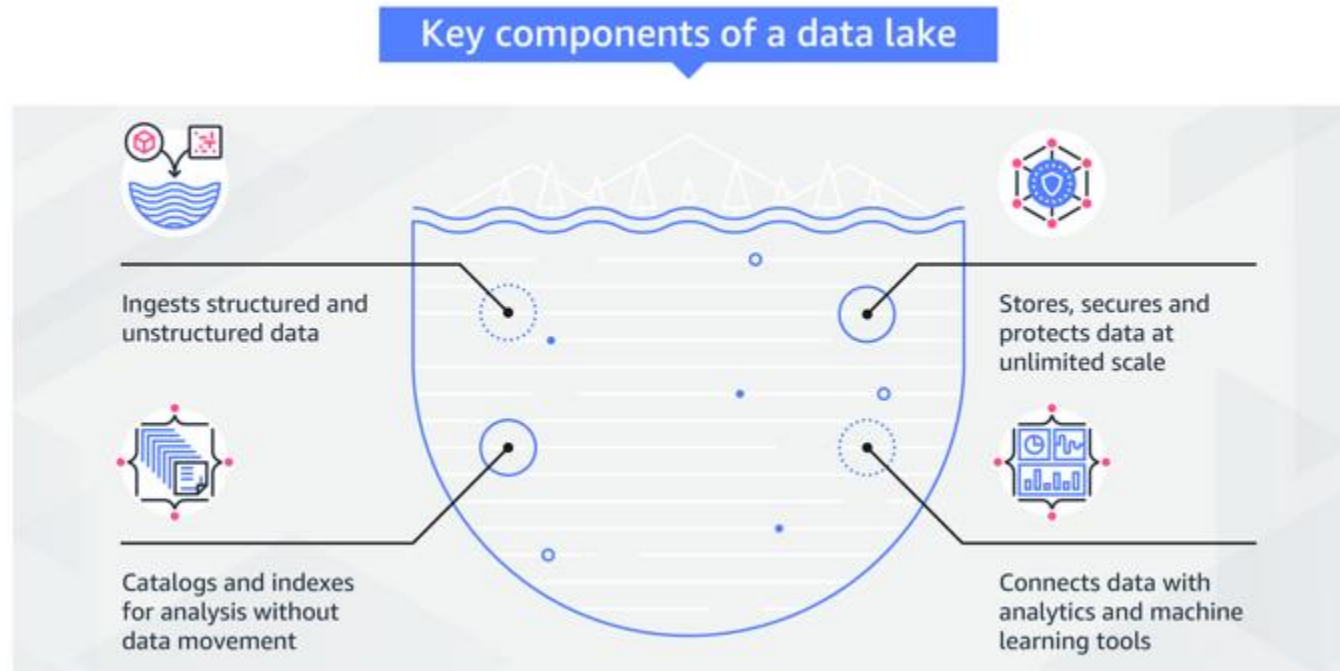
- La sua peculiarità è di consentire il recupero e l'organizzazione del dato **nel momento della fruizione** secondo il tipo di analisi che si intende effettuare
- I Data Lake forniscono una **visualizzazione non elaborata** dei dati
 - Per "dati non elaborati" si intendono quei dati che non sono ancora stati elaborati per uno scopo specifico: un dato in un Data Lake non viene definito fino al momento in cui non viene eseguita una query che lo coinvolga
- I **data scientist** possono accedere ai dati non elaborati mentre utilizzano strumenti di analisi avanzati o di modellazione predittiva



Data Lake

Il termine Data Lake è stato introdotto da James Dixon, Chief Technology Officer di Pentaho, ha coniato il termine per metterlo in contrasto con il **data mart**, un archivio di attributi interessanti derivati da dati grezzi, ha sostenuto che i data mart hanno diversi problemi intrinseci, come il silo di informazioni

- «*Lago di dati*»: l'immagine dell'acqua in questo caso è calzante, perché questo repository gestisce un pool di dati al suo stato naturale, come se fossero forme fluide non ancora filtrate o suddivise in pacchetti
- «*Schema on read*»: i dati vengono elaborati solamente nel momento in cui sono pronti per essere utilizzati

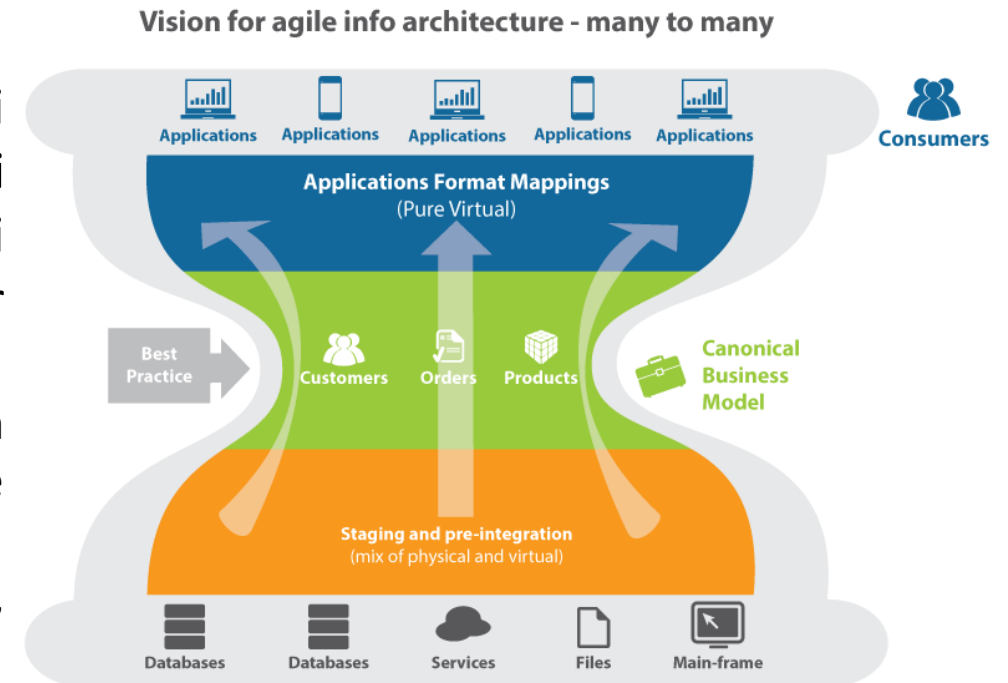


Architettura di un Data Lake

Un Data Lake ha un'**architettura piatta** in cui i dati possono essere non strutturati, semi-strutturati o strutturati e raccolti da diverse fonti: una volta inseriti nel Data Lake, i dati devono essere contrassegnati con **metadati**

I data scientist possono accedere a tutti i dati, analizzarli sfruttando gli strumenti di analisi dei big data e di machine learning, condividerli e fare riferimenti incrociati, anche tra dati eterogenei da campi diversi, per **ottenere nuove informazioni**

- Deve disporre di una **governance** che assicuri una manutenzione continua per rendere i dati fruibili e accessibili
- Deve essere dotato di una struttura di indici centralizzata, che copra dati e **metadati**, comprese informazioni su fonti, *versioning*, veridicità e livello di accuratezza



Differenze tra Data Lake e Data Warehouse

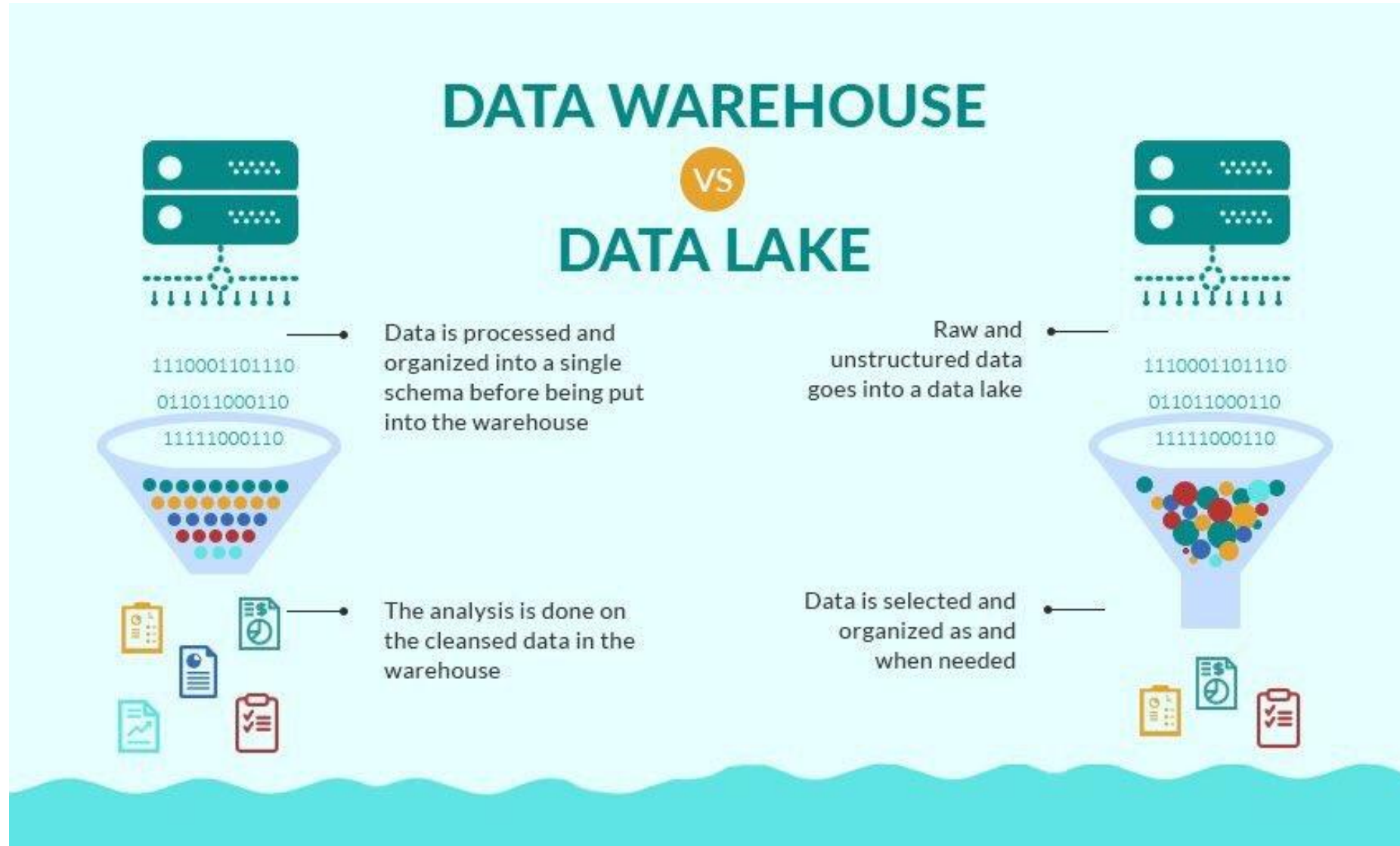
Data Lake e Data Warehouse sono due concetti diversi e servono a scopi profondamente distinti

Il Data Lake fornisce una base dati analizzabile in tempo reale da data scientist con obiettivi di analisi diversi

Il Data Warehouse produce report standardizzati e che prevedono Etl e processi di trasformazione complessi

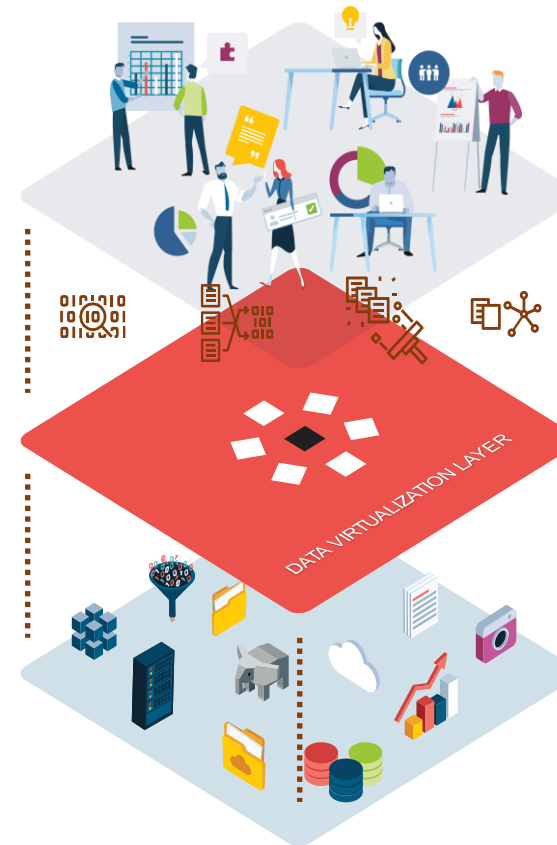
	Data Warehouse	Data Lake
Scopo	rendere disponibile una visione di dati elaborati per un processo ben preciso	rendere disponibile una visione dei dati a supporto delle attività di data discovery
Modello dei dati	modello di dati strutturato e progettato per la reportistica, struttura definita a priori, dati scritti nella struttura predefinita e poi letti nel formato desiderato (Schema-on-write)	archivia dati non strutturati, non elaborati e senza uno scopo predefinito, acquisiti nel formato nativo, ogni elemento riceve un identificatore e un insieme di metadati a corredo (Schema-on-read)
Raccolta dei dati	il processo per l'inserimento dei dati è caratterizzato da una fase preliminare di standardizzazione delle informazioni e di modellazione dei dati tramite processi di ETL (Extraction, Transformation e Loading), è un processo complesso che richiede tempo	non necessita di una strutturazione ex ante del dato, può accogliere dati strutturati, semi-strutturati e destrutturati, dati con formati molto differenti senza necessità di doverli uniformare e "normalizzare", si può iniziare a raccogliere i dati fin da subito e decidere come utilizzarli in un secondo momento
Agilità e flessibilità	cambiare la struttura può risultare molto dispendioso in termini di tempo	consente di configurare e riconfigurare facilmente modelli, query e app live e di procedere al Data Analytics in modo flessibile
Utilizzo dei dati	con la struttura predefinita viene utilizzato dagli analisti e dagli utenti aziendali che sanno in anticipo di quali dati hanno bisogno per la reportistica standard	viene utilizzato soprattutto dai data scientist e dagli analisti che effettuano ricerche «libere», applicando di volta in volta filtri e analisi più avanzati

Differenze tra Data Lake e Data Warehouse



Vantaggi del Data Lake

- Nei sistemi tradizionali è necessario prevedere in anticipo tutti gli usi dei dati di cui si avrà bisogno: nel Data Warehouse modificare o aumentare la struttura del database implica tempi e costi
- Con il mutare delle esigenze, **cambiano** anche i **requisiti di analisi**: professionisti diversi in azienda hanno bisogno di diversi set di dati
- Il Data Lake supera il problema della struttura del database ed i costi di consolidamento dei dati: **l'accesso alle informazioni è sempre immediato e real-time**, gli *insight* ottenuti diventano accessibili a chiunque abbia i permessi tramite una vista unificata dei dati



Connessione

1

Creazioni di viste «normalizzate» a partire da qualsiasi sorgente dati

Modellazione

2

Esplorazione, trasformazione, preparazione e miglioramento della qualità, integrazione

Consumo

3

Condivisione, consegna, pubblicazione, governo e collaborazione

In-memory

Elaborazione in-memory (IMC): archivia i dati nella **RAM** invece che nel database in hosting su dischi

Elimina i requisiti delle transazioni I/O e ACID delle applicazioni OLTP e **accelera esponenzialmente le velocità di accesso ai dati**, poiché le informazioni archiviate nella RAM sono disponibili istantaneamente, mentre i dati conservati sui dischi sono soggetti ai limiti imposti dalle velocità di rete e disco

- L'IMC consente un'elaborazione estremamente rapida, potenzia l'elaborazione degli eventi complessi, accelera la generazione di report e rende più rapido e preciso il processo decisionale, migliora l'esperienza utente e incrementa la soddisfazione del cliente

In-memory database (IMDB), o main memory database system ("sistema di basi di dati in memoria centrale", MMDB), o memory-resident database ("base di dati residente in memoria"): è un DBMS che gestisce i dati nella memoria centrale, diversamente dai DBMS che mantengono i dati su memorie di massa (dischi rigidi), garantendo velocità molto più alte rispetto ai DBMS su memorie di massa

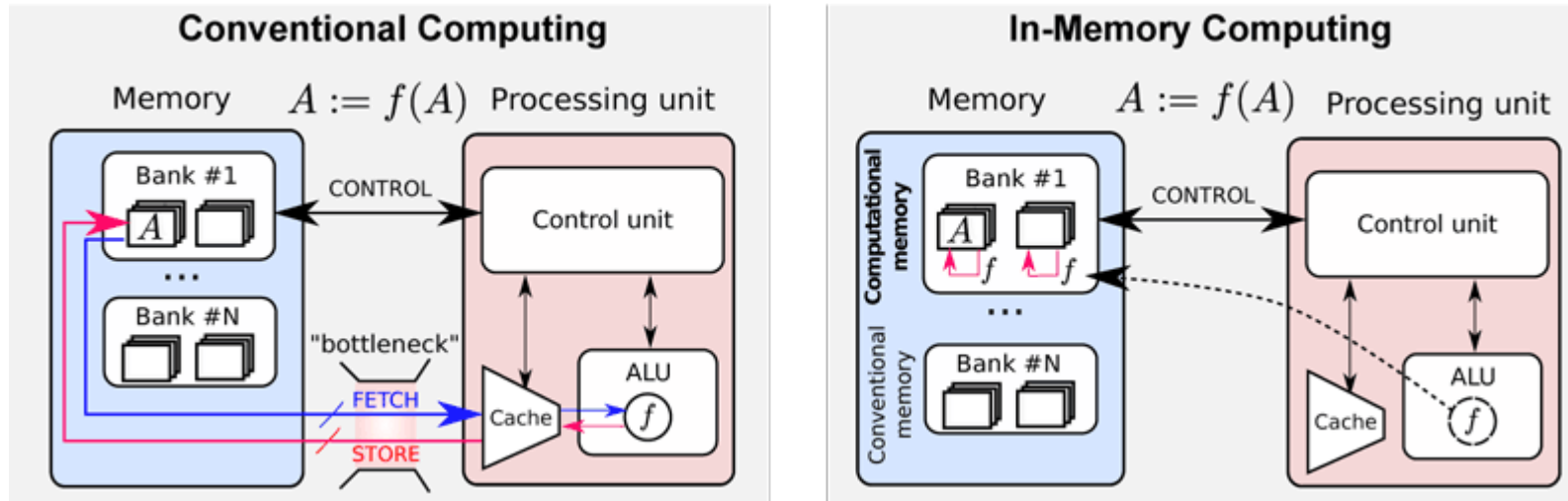
- L'elaborazione in memory può accelerare il database di un fattore che può arrivare a 100 volte
- Un IMDB può essere implementato anche con **strutture** differenti da quelle utilizzate per l'approccio relazionale (tabelle), quali quelle suggerite dal **modello reticolare** (puntatori), dal **modello gerarchico** (alberi) o dal **modello a oggetti** (oggetti complessi e nidificati)

<https://www.hpe.com/it/it/what-is/in-memory-computing.html>

https://it.wikipedia.org/wiki/In-memory_database

<https://www.zerounoweb.it/techtarget/searchdatacenter/cosa-ce-da-sapere-sullelaborazione-in-memory/>

In-memory vs Conventional Computing



Confronto tra un'architettura di elaborazione convenzionale (a sinistra) e l'elaborazione in memoria (a destra). Adattato da A. Sebastian et al. : "Temporal correlation detection using computational phase-change memory", Nature Communications 8, 1115, 2017

<https://ercim-news.ercim.eu/en115/r-i/2115-in-memory-computing-towards-energy-efficient-artificial-intelligence>

<https://www.nature.com/articles/s41467-017-01481-9>

http://www.erc-projstor.eu/resources/IBM_Preprint_Y2017_Sebastian_NatComm_merged.pdf

Quest'opera è stata rilasciata con licenza Creative Commons Attribuzione - Non commerciale - Condividi allo stesso modo 4.0 Internazionale.
Per leggere una copia della licenza visita il sito web <http://creativecommons.org/licenses/by-nc-sa/4.0/>.